A Bayes factor with reasonable model selection consistency for ANOVA model

Yuzo Maruyama

The University of Tokyo
e-mail: maruyama@csis.u-tokyo.ac.jp

Abstract: For the balanced ANOVA setup, we propose a new closed form Bayes factor without integral representation, which is however based on fully Bayes method, with reasonable model selection consistency for two asymptotic situations (either number of levels of the factor or number of replication in each level goes to infinity). Exact analytical calculation of the marginal density under a special choice of the priors enables such a Bayes factor.

AMS 2000 subject classifications: Primary 62F15, 62F07; secondary 62A10.
Keywords and phrases: Bayesian model selection, model selection consistency, fully Bayes method, Bayes factor.

1. Introduction

We start with a simple one-way balanced ANalysis-Of-VAriance (ANOVA). There are two possible models. In one model, all random variables have the same mean. In the other model, random variables in each level has a different mean. Formally, the independent observations y_{ij} $(i = 1, \ldots, p,\ j = 1, \ldots, r,\ n = pr)$ are assumed to arise from the linear model:

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2) \tag{1.1}$$

where μ, α_i $(i = 1, \ldots, p)$ and σ^2 are unknown. We assume $\sum \alpha_i = 0$ as uniqueness constraint. Clearly two models are written as follows:

$$\begin{aligned} \mathcal{M}_1 :\ & \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)' = \mathbf{0} \\ \text{vs } \mathcal{M}_{A+1} :\ & \boldsymbol{\alpha} \in \{\boldsymbol{a} \in \mathcal{R}^p | \boldsymbol{a} \neq \mathbf{0},\ \boldsymbol{a}'\mathbf{1}_p = 0\}. \end{aligned} \tag{1.2}$$

In (1.2), A means the name of the factor and the subscript $A + 1$ is from the fact that $E[y_{ij}]$ in (1.1) consists of the sum of the constant term and the level of the factor.

In this paper, we will consider Bayesian model selection based on Bayes factor for ANOVA problem. Model comparison, which refers to using the data in order to decide on the plausibility of two or more competing models, is a common problem in modern statistical science. In the Bayesian framework, the approach for model selection and hypothesis testing is essentially same, whereas there is a big difference in classical frequentist procedures for model selection and

hypothesis testing. A natural approach is to use Bayes factor (ratio of marginal densities of two models), which is based on the posterior model probabilities (Kass & Raftery (1995)). That is the reason why we take Bayesian approach based on Bayes factor in this paper.

One of the most important topic on Bayesian model selection is consistency. Consistency means that the true model will be chosen if enough data are observed, assuming that one of the competing models is true. It is well-known that BIC by Schwarz (1978) has consistency in classical (so called "$n > p$") situation. As a variant of "$p > n$" problem, which is hot in modern statistics, the consistency in the case where $p \to \infty$ and r is fixed in one-way ANOVA setup, has been considered by Stone (1979) and Berger et al. (2003). In the following, "CASE I" and "CASE II" denote the cases where

I. r goes to infinity and p is fixed,
II. p goes to infinity and r is fixed,

respectively. Under known σ^2 and CASE II, Stone (1979) showed that BIC always chooses the null model $\mathcal{M}_1$ (that is, BIC is not consistent under $\mathcal{M}_{A+1}$) even if $\boldsymbol{\alpha}'\boldsymbol{\alpha}/\{p\sigma^2\}$ is sufficiently large. This is reasonable because BIC is originally derived by the Laplace approximation under classical situation. Under known σ^2, Berger et al. (2003) proposed the Bayesian criterion called GBIC, which is derived by the Laplace approximation under CASE II. Then they showed that GBIC has model selection consistency under CASE II.

Generally, the original representation of Bayes factors or marginal densities involve integral. In the normal linear model setup, even if conjugate prior is used, hyperparameter and its prior distribution are usually introduced in order to guarantee objectivity, which is called fully Bayes method. (On the other hand, in empirical Bayes method, maximization of the conditional marginal density given hyperparameter with respect to hyperparameter is applied.) Since finding a prior of hyperparameter, which enables analytical calculation completely, is considered as extremely hard, the Laplace approximation has been applied. Needless to say, the Laplace approximation needs some assumptions, in particular, on "what goes to infinity". However, when both p and r are large (or small) in analysis of real data, the answer to the question which type of the Laplace approximation is more appropriate, is obscure. Moreover an approximated Bayes factor under one assumption does not necessarily have consistency on the other assumption, which is not good for practitioners. Therefore Bayes factor

1. without integral representation, which is however based on fully Bayes method,
2. with model selection consistency for two asymptotic situations, CASE I and II

is desirable, which we will propose in this paper. Actually, a special choice of the prior of hyperparameter, which completely enables analytical calculation of the marginal density, is the key in the paper.

Eventually the Bayes factor which we recommend is given by

$$\mathrm{BF}_{FB}[\mathcal{M}_{A+1};\mathcal{M}_1]=\frac{\Gamma(p/2)\Gamma(p(r-1)/2)}{\Gamma(1/2)\Gamma(\{pr-1\}/2)}\left(\frac{W_E}{W_T}\right)^{-p(r-1)/2+1/2}, \tag{1.3}$$

where

$$\frac{W_E}{W_T}=\frac{\sum_{ij}(y_{ij}-\bar{y}_{i\cdot})^2}{\sum_{ij}(y_{ij}-\bar{y}_{\cdot\cdot})^2},\ \bar{y}_{i\cdot}=\frac{\sum_j y_{ij}}{r},\ \bar{y}_{\cdot\cdot}=\frac{\sum_{ij} y_{ij}}{pr}.$$

It is not only exactly proportional to the posterior probability of $\mathcal{M}_{A+1}$, but also a function of W_E/W_T, which is fundamental aggregated information of one-way ANOVA, from the frequentist viewpoint.

The rest of the paper is organized as follows. In Section 2, we review methods for Bayesian model selection based on Bayes factor, give a concrete form of the prior we use for one-way balanced ANOVA and eventually propose the Bayes factor given by (1.3). In Section 3, we show that the Bayes factor has a reasonable model selection consistency. In Section 4, we extend results in Section 2 and 3 to two-way balanced ANOVA setup.

2. A Bayes factor for one-way ANOVA

In this section, we will propose the Bayes factor for one-way ANOVA, which is given by (1.3). In Sub-section 2.1, we first re-parameterize the ANOVA model given by (1.1) and then give priors for it. In Sub-section 2.2, we derive marginal densities under $\mathcal{M}_1$ and $\mathcal{M}_{A+1}$ and eventually the Bayes factor using them.

2.1. re-parameterized ANOVA and priors

Let

$$\boldsymbol{y}=(y_{11},\ldots,y_{1r},y_{21},\ldots,y_{2r},\ldots,y_{p1},\ldots,y_{pr})' \text{ and } \boldsymbol{X}=\boldsymbol{I}_p\otimes\mathbf{1}_r.$$

Then the linear model given by (1.1) is written as

$$\boldsymbol{y}=\mu\mathbf{1}_n+\boldsymbol{X}\boldsymbol{\alpha}+\boldsymbol{\epsilon}.$$

Further the centered matrix of $\boldsymbol{X}$, $\boldsymbol{X}-p^{-1}\mathbf{1}_n\mathbf{1}_p'$, is decomposed by the singular value decomposition as

$$\sqrt{r}\boldsymbol{U}\boldsymbol{W}'=\sqrt{r}(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_{p-1})(\boldsymbol{w}_1,\ldots,\boldsymbol{w}_{p-1})'$$

where $\boldsymbol{U}$ and $\boldsymbol{W}$ are $n\times(p-1)$ and $p\times(p-1)$ orthogonal matrices, respectively. Let $\boldsymbol{\theta}=\sqrt{r}\boldsymbol{W}'\boldsymbol{\alpha}\in\mathcal{R}^{p-1}$. Then the one-way ANOVA is re-parameterized as the linear regression model

$$\boldsymbol{y}=\mu\mathbf{1}_n+\boldsymbol{U}\boldsymbol{\theta}+\boldsymbol{\epsilon} \tag{2.1}$$

and two models are written as $\mathcal{M}_1$: $\boldsymbol{\theta}=\mathbf{0}_{p-1}$ and $\mathcal{M}_{A+1}$: $\boldsymbol{\theta}\in\mathcal{R}^{p-1}\setminus\{\mathbf{0}_{p-1}\}$.

In model selection in the Bayesian framework, the specification of priors are needed on the models and parameters in each model. For the former, let $\Pr(\mathcal{M}_1) = \Pr(\mathcal{M}_{A+1}) = 1/2$ as usual. For the latter, at moment, we just write joint prior densities as $p(\mu, \sigma^2)$ for $\mathcal{M}_1$ and $p(\mu, \boldsymbol{\theta}, \sigma^2)$ for $\mathcal{M}_{A+1}$. From the Bayes theorem, $\mathcal{M}_{A+1}$ is chosen when $\Pr(\mathcal{M}_{A+1}|\boldsymbol{y}) > 1/2$ where

$$\Pr(\mathcal{M}_{A+1}|\boldsymbol{y}) = \frac{\mathrm{BF}(\mathcal{M}_{A+1}; \mathcal{M}_1)}{1 + \mathrm{BF}(\mathcal{M}_{A+1}; \mathcal{M}_1)}$$

and $\mathrm{BF}(\mathcal{M}_{A+1}; \mathcal{M}_1)$ is the Bayes factor given by

$$\mathrm{BF}(\mathcal{M}_{A+1}; \mathcal{M}_1) = m_{A+1}(\boldsymbol{y})/m_1(\boldsymbol{y}). \tag{2.2}$$

In other words, $\mathcal{M}_{A+1}$ is chosen if and only if $\mathrm{BF}(\mathcal{M}_{A+1}; \mathcal{M}_1) > 1$. In (2.2), $m_\gamma(\boldsymbol{y})$ is the marginal density under $\mathcal{M}_\gamma$ for $\gamma = 1, A+1$ as follows:

$$\begin{aligned} m_1(\boldsymbol{y}) &= \iint p(\boldsymbol{y}|\mu, \sigma^2) p(\mu, \sigma^2) d\mu d\sigma^2 \\ m_{A+1}(\boldsymbol{y}) &= \iiint p(\boldsymbol{y}|\mu, \boldsymbol{\theta}, \sigma^2) p(\mu, \boldsymbol{\theta}, \sigma^2) d\mu d\boldsymbol{\theta} d\sigma^2, \end{aligned}$$

where $p(\boldsymbol{y}|\mu, \sigma^2)$ and $p(\boldsymbol{y}|\mu, \boldsymbol{\theta}, \sigma^2)$ are sampling densities of $\boldsymbol{y}$ under $\mathcal{M}_1$ and $\mathcal{M}_{A+1}$, respectively.

In this paper, we use the following (improper) joint density

$$p(\mu, \sigma^2) = p(\mu)p(\sigma^2) = 1 \times \sigma^{-2} \tag{2.3}$$

for $\mathcal{M}_1$ and

$$p(\mu, \boldsymbol{\theta}, \sigma^2) = p(\mu)p(\sigma^2)p(\boldsymbol{\theta}|\sigma^2) = \frac{1}{\sigma^2} \int_0^\infty p(\boldsymbol{\theta}|g, \sigma^2) p(g) dg \tag{2.4}$$

for $\mathcal{M}_{A+1}$. Note that $p(\mu)p(\sigma^2) = \sigma^{-2}$ in both (2.3) and (2.4) is a popular non-informative prior. It is improper, but justified because μ and σ^2 are included in both $\mathcal{M}_1$ and $\mathcal{M}_{A+1}$. The proper densities $p(\boldsymbol{\theta}|g, \sigma^2)$ and $p(g)$ will be specified in the next subsection. Recall that we will consider fully Bayes method, which means the joint prior densities $p(\mu, \sigma^2)$ and $p(\mu, \boldsymbol{\theta}, \sigma^2)$ do not depend on the observations $\boldsymbol{y}$. On the other hand, the prior densities for empirical Bayes method do (George & Foster (2000)).

2.2. Marginal densities and the Bayes factor

First we derive the marginal density under $\mathcal{M}_1$. Using the Pythagorean relation

$$\|\boldsymbol{y} - \mu \mathbf{1}_n\|^2 = n(\bar{y}_{\cdot\cdot} - \mu)^2 + W_T,$$

where $\bar{y}_{\cdot\cdot} = n^{-1} \sum_{i,j} y_{ij}$ and $W_T = \|\boldsymbol{y} - \bar{y}_{\cdot\cdot}\mathbf{1}_n\|^2 = \sum_{ij}(y_{ij} - \bar{y}_{\cdot\cdot})^2$, we have

$$m_1(\boldsymbol{y}) = \int_{-\infty}^\infty \int_0^\infty \frac{1}{(2\pi)^{n/2}\sigma^{n+2}} \exp\left(-\frac{\|\boldsymbol{y} - \mu\mathbf{1}_n\|^2}{2\sigma^2}\right) d\mu d\sigma^2$$

$$= \frac{n^{1/2}}{(2\pi)^{n/2-1/2}} \int_0^\infty \frac{1}{\sigma^{n+1}} \exp\left(-\frac{W_T}{2\sigma^2}\right) d\sigma^2$$
$$= \frac{n^{1/2}\Gamma(\{n-1\}/2)}{\pi^{(n-1)/2}} \{W_T\}^{-(n-1)/2}.$$

Notice that W_T is called "total sum of squares" in the ANOVA context. The total sum of squares, W_T, is identically partitioned as the sum of "within group sum of squares" W_E and "between group sum of squares" W_H as follows:

$$\begin{aligned} W_T &= \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2 + \|\hat{\boldsymbol{y}} - \bar{y}_{\cdot\cdot}\mathbf{1}_n\|^2 \\ &= \textstyle\sum_{ij}(y_{ij} - \bar{y}_{i\cdot})^2 + \sum_{ij}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 \\ &= W_E + W_H, \end{aligned}$$

where $\bar{y}_{i\cdot} = r^{-1}\sum_j y_{ij}$ for $i = 1, \ldots, p$ and $\hat{\boldsymbol{y}} = (\bar{y}_{1\cdot}, \ldots \bar{y}_{p\cdot})' \otimes \mathbf{1}_r$.

Then we derive the marginal density under $\mathcal{M}_{A+1}$. As explained in the previous subsection, (2.4) is the form of the joint density we use in this paper. As $p(\boldsymbol{\theta}|g, \sigma^2)$, we use so-called Zellner (1986)'s g-prior

$$p(\boldsymbol{\theta}|\sigma^2, g) = N_{p-1}(\mathbf{0}, g\sigma^2(\boldsymbol{U}'\boldsymbol{U})^{-1}) = N_{p-1}(\mathbf{0}, g\sigma^2\boldsymbol{I}_{p-1}).$$

There are many papers which use g-priors including George & Foster (2000), Liang et al. (2008), Maruyama & George (2008) in Bayesian model selection context. In the Stein estimation context, this type of shrinkage priors is known as Strawderman (1971)'s prior. See Maruyama & Strawderman (2005) for the detail.

Using the relationship

$$\begin{aligned} &\|\boldsymbol{y} - \mu\mathbf{1}_n - \boldsymbol{U}\boldsymbol{\theta}\|^2 + g^{-1}\|\boldsymbol{\theta}\|^2 \\ &= n(\bar{y}_{\cdot\cdot} - \mu)^2 + \|\boldsymbol{y} - \bar{y}_{\cdot\cdot}\mathbf{1}_n - \boldsymbol{U}\boldsymbol{\theta}\|^2 + g^{-1}\|\boldsymbol{\theta}\|^2 \\ &= n(\bar{y}_{\cdot\cdot} - \mu)^2 + \frac{g+1}{g}\left\|\boldsymbol{\theta} - \frac{g\boldsymbol{U}'(\boldsymbol{y} - \bar{y}_{\cdot\cdot}\mathbf{1}_n)}{g+1}\right\|^2 + \frac{W_T + gW_E}{g+1}, \end{aligned}$$

we have the conditional marginal density of $\boldsymbol{y}$ given g under $\mathcal{M}_{A+1}$,

$$\begin{aligned} m_{A+1}(\boldsymbol{y}|g) &= \int_{-\infty}^{\infty}\int_{\mathcal{R}^{p-1}}\int_0^\infty p(\boldsymbol{y}|\mu, \boldsymbol{\theta}, \sigma^2)p(\boldsymbol{\theta}|\sigma^2, g)p(\sigma^2)d\mu d\boldsymbol{\theta} d\sigma^2 \\ &= \int_{-\infty}^{\infty}\int_{\mathcal{R}^{p-1}}\int_0^\infty \frac{1}{(2\pi\sigma^2)^{n/2}}\frac{1}{(2\pi g\sigma^2)^{(p-1)/2}} \\ &\qquad \times \exp\left(-\frac{\|\boldsymbol{y} - \mu\mathbf{1}_n - \boldsymbol{U}\boldsymbol{\theta}\|^2}{2\sigma^2} - \frac{\|\boldsymbol{\theta}\|^2}{2g\sigma^2}\right) p(\sigma^2)d\mu d\boldsymbol{\theta} d\sigma^2 \\ &= \int_0^\infty \frac{n^{1/2}(1+g)^{-(p-1)/2}}{(2\pi\sigma^2)^{(n-1)/2}} \exp\left(-\frac{W_T + gW_E}{2\sigma^2(g+1)}\right)\frac{1}{\sigma^2}d\sigma^2 \\ &= m_1(\boldsymbol{y})\frac{(1+g)^{(n-p)/2}}{(g\{W_E/W_T\} + 1)^{(n-1)/2}}. \end{aligned}$$

The (fully) marginal density is given by

$$m_{A+1}(\boldsymbol{y}) = \int_0^\infty m_{A+1}(\boldsymbol{y}|g)p(g)dg$$

which is usually calculated by numerical methods like MCMC or by approximation like Laplace method. But, in this paper, very nice analytical results will be derived by choosing a following special prior of g.

As the prior of g, we use Pearson Type VI or beta-prime distribution

$$p(g) = \{B(a+1, b+1)\}^{-1} g^b (1+g)^{-a-b-2}, \tag{2.5}$$

which is clearly proper if $a > -1$ and $b > -1$. In particular, when $b = (n-p)/2 - a - 2$, we get a closed simple form of the marginal density

$$\begin{aligned} m_{A+1}(\boldsymbol{y}) &= m_1(\boldsymbol{y}) \int_0^\infty \frac{(1+g)^{(n-p)/2}}{(g\{W_E/W_T\}+1)^{(n-1)/2}} p(g)dg \\ &= m_1(\boldsymbol{y}) \frac{\Gamma(p/2+a+1/2)\Gamma((n-p)/2)}{\Gamma(a+1)\Gamma(\{n-1\}/2)} \left(\frac{W_E}{W_T}\right)^{-(n-p-2)/2+a}. \end{aligned}$$

If $b \neq (n-p)/2-a-2$, there remains an integral including W_E/W_T on $m_{A+1}(\boldsymbol{y})$. Actually "hyper-g priors" given by Liang et al. (2008) corresponds to the case $b = 0$. They had no other choice to use the Laplace approximation.

For the choice of a, my recommendation is $a = -1/2$. We will describe it briefly. The asymptotic behavior of $p(g)$ given by (2.5), for sufficiently large g, is proportional to g^{-a-2}. From the Tauberian Theorem, which is well-known for describing the asymptotic behavior of the Laplace transform, we have

$$p(\boldsymbol{\theta}|\sigma^2) = \int_0^\infty p(\boldsymbol{\theta}|\sigma^2, g)p(g)dg \approx (\sigma^2)^{a+1} \|\boldsymbol{\theta}\|^{-(p+2a+1)}, \tag{2.6}$$

for sufficiently large $\boldsymbol{\theta} \in \mathcal{R}^{p-1}$, $a > -1$ and $b > -1$. Hence the asymptotic tail behavior of $p(\boldsymbol{\theta}|\sigma^2)$ for $a = -1/2$ is multivariate Cauchy, $\|\boldsymbol{\theta}\|^{-(p-1)-1}$, which has been recommended by Zellner & Siow (1980) and others in objective Bayes context.

Finally the Bayes factor which we recommend is written as

$$\begin{aligned} \mathrm{BF}_{FB}[\mathcal{M}_{A+1}; \mathcal{M}_1] &= \frac{m_{A+1}(\boldsymbol{y})}{m_1(\boldsymbol{y})} \\ &= \frac{\Gamma(p/2)\Gamma(p(r-1)/2)}{\Gamma(1/2)\Gamma(\{pr-1\}/2)} \left(\frac{W_E}{W_T}\right)^{-p(r-1)/2+1/2}, \end{aligned} \tag{2.7}$$

where the subscript FB means "Fully Bayes". It is not only exactly proportional to the posterior probability $\Pr(\mathcal{M}_{A+1}|\boldsymbol{y})$, but also a function of W_T and W_E only through W_E/W_T, which is fundamental aggregated information of one-way ANOVA, from the frequentist viewpoint.

Remark 2.1. The most advantage of BF_{FB} over existing Bayes factors is its excellent closed form. Many Bayes factors based on fully Bayes method including intrinsic Bayes factor by Casella et al. (2009) have the closed forms. Since it usually involves the integral in the representation, they have to apply the Laplace approximation in practice. For example, Casella et al. (2009) and Moreno et al. (2009) applied different types of the Laplace approximation to the same Bayes factor with integral representation. However, the answer to the question which type of the Laplace approximation is more appropriate, is obscure for some cases (for example, in the case where both p and r are large (or small) in ANOVA problem). On the other hand, BF_{FB} does not require thought and has a reasonable model selection consistency for two cases ($r \to \infty$, fixed p) and ($p \to \infty$, fixed r), as seen in the next section.

3. Model selection consistency

In this section, we consider the model selection consistency in the case where $n = pr$ approaches the infinity. In concrete, we consider two cases ($r \to \infty$, fixed p) and ($p \to \infty$, fixed r). Generally, the posterior consistency for model choice is defined as

$$\operatorname*{plim}_{n\to\infty} \Pr(\mathcal{M}_\gamma|y) = 1 \tag{3.1}$$

or equivalently

$$\operatorname*{plim}_{n\to\infty} \mathrm{BF}[\mathcal{M}_\gamma; \mathcal{M}_{\gamma'}] = \infty$$

when $\mathcal{M}_\gamma$ is the true model and $\mathcal{M}_{\gamma'}$ is not. Here plim denotes convergence in probability and the probability distribution in (3.1) is the sampling distribution under the true model $\mathcal{M}_\gamma$. We will show that $\mathrm{BF}_{FB}[\mathcal{M}_{A+1}; \mathcal{M}_1]$ given by (2.7) has a reasonable model selection consistency. As the competitor of BF_{FB}, the Bayes factor based on BIC

$$\mathrm{BF}_{BIC}[\mathcal{M}_{A+1}; \mathcal{M}_1] = \left(\frac{W_E}{W_T}\right)^{-pr/2} (pr)^{-(p-1)/2}$$

will be considered.

Theorem 3.1. *1. Assume $r \to \infty$ and p is fixed.*

(a) BF_{BIC} and BF_{FB} are consistent whichever the true model is.

2. Assume $p \to \infty$ and r is fixed. Also let $c_A = \lim_{p\to\infty} \sum_{i=1}^{p} \alpha_i^2/\{p\sigma^2\}$.

(a) BF_{BIC} and BF_{FB} are consistent under $\mathcal{M}_1$.

(b) BF_{FB} is consistent [inconsistent] under $\mathcal{M}_{A+1}$ when $c_A > [<] h(r)$ where

$$h(r) = r^{1/(r-1)} - 1. \tag{3.2}$$

(c) BF_{BIC} is inconsistent under $\mathcal{M}_{A+1}$ for any $c_A > 0$.

Note: $h(r)$ is clearly a convex decreasing function in r which satisfies $h(2) = 1$, $h(5) \doteq 0.5$, $h(10) \doteq 0.29$, and $h(\infty) = 0$.

Proof. In the proof, $\xrightarrow{\mathrm{P}}$ denotes convergence in probability. Note

$$\frac{W_E}{W_T} = \frac{W_E/\sigma^2}{W_T/\sigma^2} \sim \frac{\chi^2_{pr-p}}{\chi^2_{pr-p} + \chi^2_{p-1}[r\sum \alpha_i^2/\sigma^2]}.$$

Hence we have

$$\frac{W_E}{W_T} \xrightarrow{\mathrm{P}} \begin{cases} \left(1 + \chi^2_{p-1}/pr\right)^{-1} & \text{under } \mathcal{M}_1 \\ \left(1 + \sum \alpha_i^2/\{p\sigma^2\}\right)^{-1} & \text{under } \mathcal{M}_{A+1} \end{cases} \tag{3.3}$$

when $r \to \infty$ and p is fixed and

$$\frac{W_E}{W_T} \xrightarrow{\mathrm{P}} \begin{cases} (1 - 1/r) & \text{under } \mathcal{M}_1 \\ (1 - 1/r)/(1 + c_A) & \text{under } \mathcal{M}_{A+1} \end{cases} \tag{3.4}$$

when $p \to \infty$, r is fixed and $c_A = \lim_{p\to\infty} \sum_{i=1}^{p} \alpha_i^2/\{p\sigma^2\}$ is assumed.

For the asymptotic behavior of the gamma function, Stiring's formula,

$$\Gamma(ax + b) \approx \sqrt{2\pi} e^{-ax} (ax)^{ax+b-1/2} \tag{3.5}$$

for sufficiently large x is useful. Here $f \approx g$ means $\lim f/g = 1$. Using (3.5), we get

$$\frac{\Gamma(\{pr - p\}/2)}{\Gamma(\{pr - 1\}/2)} \approx (pr/2)^{-(p-1)/2}, \tag{3.6}$$

when $r \to \infty$ and p is fixed, and

$$\frac{\Gamma(p/2)\Gamma(\{pr - p\}/2)}{\Gamma(\{pr - 1\}/2)} \approx \frac{\sqrt{2\pi} r}{(r-1)^{1/2}} \left\{ \frac{r-1}{r^{r/(r-1)}} \right\}^{p(r-1)/2} \tag{3.7}$$

when $p \to \infty$ and r is fixed.

First, we consider the consistency in the case where $r \to \infty$ and p is fixed. Using (3.3) and (3.6), we have

$$\mathrm{BF}_{FB}[\mathcal{M}_{A+1}; \mathcal{M}_1] \xrightarrow{\mathrm{P}} \begin{cases} c_1(p) r^{-(p-1)/2} \exp(\chi^2_{p-1}/2) & \text{under } \mathcal{M}_1 \\ c_1(p) r^{-(p-1)/2} (1 + \sum \alpha_i^2/\{p\sigma^2\})^{pr} & \text{under } \mathcal{M}_{A+1}, \end{cases}$$

which goes to zero under $\mathcal{M}_1$ and infinity under $\mathcal{M}_{A+1}$ respectively, and

$$\mathrm{BF}_{BIC}[\mathcal{M}_{A+1}; \mathcal{M}_1] \xrightarrow{\mathrm{P}} \begin{cases} c_2(p) r^{-(p-1)/2} \exp(\chi^2_{p-1}/2) & \text{under } \mathcal{M}_1 \\ c_2(p) r^{-(p-1)/2} (1 + \sum \alpha_i^2/\{p\sigma^2\})^{pr} & \text{under } \mathcal{M}_{A+1}, \end{cases}$$

which goes to zero under $\mathcal{M}_1$ and infinity under $\mathcal{M}_{A+1}$ respectively. In the above, $c_1(p)$ and $c_2(p)$ are given by

$$c_1(p) = (p/2)^{-(p-1)/2}\Gamma(p/2)/\Gamma(1/2) \text{ and } \ c_2(p) = p^{-(p-1)/2}.$$

Thus part 1a of the theorem follows.

Then we consider the consistency in the case where $p \to \infty$ and r is fixed. Using (3.4) and (3.7), we have,

$$\mathrm{BF}_{FB}[\mathcal{M}_{A+1};\mathcal{M}_1] \xrightarrow{\mathrm{P}} \begin{cases} \sqrt{2}r(r-1)^{-1/2}r^{-p/2} & \text{under } \mathcal{M}_1 \\ \sqrt{2}r(r-1)^{-1/2}\left(\frac{1+c_A}{r^{1/(r-1)}}\right)^{p(r-1)/2} & \text{under } \mathcal{M}_{A+1} \end{cases}$$

which goes to zero under $\mathcal{M}_1$ and under $\mathcal{M}_{A+1}$ with $c_A < h(r)$. It goes to infinity under $\mathcal{M}_{A+1}$ with $c_A > h(r)$. On the other hand, we have

$$\mathrm{BF}_{BIC}[\mathcal{M}_{A+1};\mathcal{M}_1] \xrightarrow{\mathrm{P}} \begin{cases} r^{1/2}p^{-(p-1)/2}\left(\frac{r^{(r-1)}}{(r-1)^r}\right)^{p/2} & \text{under } \mathcal{M}_1 \\ r^{1/2}p^{-(p-1)/2}\left(\frac{r^{(r-1)}}{(r-1)^r}(1+c_A)^r\right)^{p/2} & \text{under } \mathcal{M}_{A+1} \end{cases}$$

which goes to zero in both models. Thus parts 2a, 2b and 2c of the theorem follow. □

Remark 3.1. We give some remarks on inconsistency shown in the theorem.

1. As shown in 2c of Theorem 3.1, BIC always chooses $\mathcal{M}_1$ when $p \to \infty$ and r is fixed, even if c_A is very large. This is interpreted as unknown variance version of Stone (1979)'s example.
2. As seen in 2b of Theorem 3.1, BF_{FB} has an inconsistency region. Actually existing such an inconsistency region has been also reported by Moreno et al. (2009). They proposed intrinsic Bayes factor for normal regression model and their upper-bound of inconsistency region for one-way ANOVA is given by

$$\frac{r-1}{(r+1)^{(r-1)/r}-1} - 1$$

   which seems to be slightly smaller than $h(r)$ given by (3.2).
3. The existence of inconsistency region for small c_A and large p is quite reasonable from the following reason. Assume new independent observations z_{ij} $(i = 1,\ldots,p,\ j = 1,\ldots,r)$ from the same model as y_{ij}. Then the difference of scaled mean squared prediction errors of $\bar{y}_{i\cdot}$ and $\bar{y}_{\cdot\cdot}$ is given by

$$\begin{aligned} \Delta[\bar{y}_{\cdot\cdot};\bar{y}_{i\cdot}] &= \frac{E_{y,z}\left[\sum_{i,j}(z_{ij}-\bar{y}_{\cdot\cdot})^2\right]}{pr\sigma^2} - \frac{E_{y,z}\left[\sum_{i,j}(z_{ij}-\bar{y}_{i\cdot})^2\right]}{pr\sigma^2} \\ &= \frac{\sum_i \alpha_i^2}{p\sigma^2} - \frac{p-1}{pr} \end{aligned}$$

for any p and r. First assume that $\mathcal{M}_1$ is true. We see that

$$\Delta[\bar{y}_{\cdot\cdot}; \bar{y}_{i\cdot}] = -(p-1)/\{pr\} < 0,$$

which is reasonable. Then assume that $\mathcal{M}_{A+1}$ is true. When $r \to \infty$ and p is fixed,

$$\lim_{r\to\infty} \Delta[\bar{y}_{\cdot\cdot}; \bar{y}_{i\cdot}] = \textstyle\sum_i \alpha_i^2/(p\sigma^2) > 0,$$

which is reasonable. On the other hand, if $p \to \infty$ and r is fixed,

$$\lim_{p\to\infty} \Delta[\bar{y}_{\cdot\cdot}; \bar{y}_{i\cdot}] = c_A - 1/r.$$

Hence when $c_A < 1/r$, $\Delta[\bar{y}_{\cdot\cdot}; \bar{y}_{i\cdot}]$ is negative even if $\mathcal{M}_{A+1}$ is true. Therefore, from the prediction point of view, the existence of inconsistency region for small c_A and large p is reasonable.

4. Berger et al. (2003) considered Bayesian model selection for one-way ANOVA. Under known variance setup, they showed that, for any prior of the form,

$$p(\boldsymbol{\alpha}) \propto \textstyle\int_0^\infty t^{p/2} \exp(-t\boldsymbol{\alpha}'\boldsymbol{\alpha}/\{2\sigma^2\}) p(t) dt,$$

with $p(t)$ having support equal to $(0, \infty)$, the Bayes factor is consistent under $\mathcal{M}_1$. They also showed that consistency under $\mathcal{M}_{A+1}$ holds if $c_A > 0$. Hence we see that their result is not extensible to the unknown σ^2 case since our beta-prime prior given by (2.5) has support equal to $(0, \infty)$. Moreno et al. (2009) discussed Berger's result from the different point of view.

Table 1 shows frequency of choice of the true model in some cases in numerical experiment. We see that it clearly guarantees the validness of Theorem 3.1.

4. two-way balanced ANOVA

In this section, we extend the results in Section 2 and 3 to two-way balanced ANOVA problem. We have n independent normal random variables y_{ijk} ($i = 1, \ldots, p,\ j = 1, \ldots, q,\ k = 1, \ldots, r,\ n = pqr$) where

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},\ \epsilon_{ijk} \sim N(0, \sigma^2).$$

We assume $\sum_i \alpha_i = \sum_j \beta_j = \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0$ as uniqueness constraint. In the two-way ANOVA, the following five sub-models are important.

$$\mathcal{M}_1 :\ E[y_{ijk}] = \mu, \quad \mathcal{M}_{A+1} :\ E[y_{ijk}] = \mu + \alpha_i, \quad \mathcal{M}_{B+1} :\ E[y_{ijk}] = \mu + \beta_j,$$
$$\mathcal{M}_{A+B+1} :\ E[y_{ijk}] = \mu + \alpha_i + \beta_j, \quad \mathcal{M}_{(A+1)(B+1)} :\ E[y_{ijk}] = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}.$$

As the one-way ANOVA is re-parameterized like (2.1), each model among $\mathcal{M}_{A+1}$, $\mathcal{M}_{B+1}$, $\mathcal{M}_{A+B+1}$, and $\mathcal{M}_{(A+1)(B+1)}$ can be also re-parameterized as follows:

$$\boldsymbol{y} = \mu \mathbf{1}_n + \boldsymbol{U}\boldsymbol{\theta} + \boldsymbol{\epsilon} \tag{4.1}$$

TABLE 1
Frequency of choice of the true model

	$p\backslash r$	2	5	10	50	100	2	5	10	50	100
		under $\mathcal{M}_1$					$c_A = 0.1$ under $\mathcal{M}_{A+1}$				
FB	2	0.77	0.89	0.94	0.98	0.98	0.28	0.26	0.30	0.80	0.97
	5	0.93	0.99	1.00	1.00	1.00	0.19	0.09	0.15	0.78	1.00
	10	0.93	0.99	1.00	1.00	1.00	0.08	0.03	0.04	0.79	1.00
	50	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.91	1.00
	100	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.97	1.00
BIC	2	0.53	0.79	0.91	0.97	0.97	0.52	0.37	0.39	0.83	0.98
	5	0.75	0.97	0.99	1.00	1.00	0.33	0.10	0.12	0.71	0.99
	10	0.94	1.00	1.00	1.00	1.00	0.07	0.01	0.01	0.50	0.99
	50	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.99
	100	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.99
		$c_A = 0.5$ under $\mathcal{M}_{A+1}$					$c_A = 1$ under $\mathcal{M}_{A+1}$				
FB	2	0.48	0.66	0.86	1.00	1.00	0.68	0.88	1.00	1.00	1.00
	5	0.39	0.59	0.90	1.00	1.00	0.59	0.93	1.00	1.00	1.00
	10	0.29	0.56	0.95	1.00	1.00	0.59	0.97	1.00	1.00	1.00
	50	0.07	0.55	1.00	1.00	1.00	0.52	1.00	1.00	1.00	1.00
	100	0.03	0.55	1.00	1.00	1.00	0.54	1.00	1.00	1.00	1.00
BIC	2	0.74	0.79	0.92	1.00	1.00	0.87	0.95	1.00	1.00	1.00
	5	0.57	0.61	0.88	1.00	1.00	0.79	0.94	1.00	1.00	1.00
	10	0.26	0.29	0.78	1.00	1.00	0.56	0.87	1.00	1.00	1.00
	50	0.00	0.00	0.03	1.00	1.00	0.00	0.03	1.00	1.00	1.00
	100	0.00	0.00	0.00	1.00	1.00	0.00	0.00	1.00	1.00	1.00
		$c_A = 2$ under $\mathcal{M}_{A+1}$					$c_A = 5$ under $\mathcal{M}_{A+1}$				
FB	2	0.85	0.99	1.00	1.00	1.00	0.98	1.00	1.00	1.00	1.00
	5	0.87	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	10	0.89	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	50	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	100	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
BIC	2	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	5	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	10	0.88	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	50	0.02	1.00	1.00	1.00	1.00	0.89	1.00	1.00	1.00	1.00
	100	0.00	0.92	1.00	1.00	1.00	0.18	1.00	1.00	1.00	1.00

where $\boldsymbol{U}$ is a $n \times s$ orthogonal matrix, $\boldsymbol{\theta} \in \mathcal{R}^s$ and

$$s = \begin{cases} p-1 & \text{under } \mathcal{M}_{A+1} \\ q-1 & \text{under } \mathcal{M}_{B+1} \\ p+q-2 & \text{under } \mathcal{M}_{A+B+1} \\ pq-1 & \text{under } \mathcal{M}_{(A+1)(B+1)}. \end{cases}$$

When $\mathcal{M}_\gamma$ for $\gamma = A+1, B+1, A+B+1$, and $(A+1)(B+1)$ and $\mathcal{M}_1$ are pairwisely compared, the corresponding Bayes factors can be derived in the same way as in Section 2.

$$\begin{aligned}
\mathrm{BF}_{FB}[\mathcal{M}_A;\mathcal{M}_1] &= \frac{\Gamma(p/2)\Gamma(p(qr-1)/2)}{\Gamma(1/2)\Gamma(\{pqr-1\}/2)}\left(1-\frac{W_A}{W_T}\right)^{-p(qr-1)/2+1/2} \\
\mathrm{BF}_{FB}[\mathcal{M}_B;\mathcal{M}_1] &= \frac{\Gamma(q/2)\Gamma(q(pr-1)/2)}{\Gamma(1/2)\Gamma(\{pqr-1\}/2)}\left(1-\frac{W_B}{W_T}\right)^{-q(pr-1)/2+1/2} \\
\mathrm{BF}_{FB}[\mathcal{M}_{A+B+1};\mathcal{M}_1] &= \frac{\Gamma(\frac{p+q-1}{2})\Gamma(\frac{pqr-p-q+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{pqr-1}{2})}\left(1-\frac{W_A+W_B}{W_T}\right)^{\frac{-pqr+p+q}{2}} \\
\mathrm{BF}_{FB}[\mathcal{M}_{(A+1)(B+1)};\mathcal{M}_1] &= \frac{\Gamma(pq/2)\Gamma(pq(r-1)/2)}{\Gamma(1/2)\Gamma(\{pqr-1\}/2)}\left(\frac{W_E}{W_T}\right)^{-pq(r-1)/2+1/2}.
\end{aligned}$$

In these expressions, W_T, W_A, W_B, W_{AB}, and W_E are sums of squares for balanced two-way ANOVA which identically satisfy

$$W_T = W_A + W_B + W_{AB} + W_E$$

and each sums of squares are defined as follows.

$$\begin{aligned}
W_T = \sum_{ijk}(y_{ijk}-\bar{y}_{\cdots})^2,\ W_A = \sum_{ijk}(\bar{y}_{i\cdot\cdot}-\bar{y}_{\cdots})^2, \\
W_B = \sum_{ijk}(\bar{y}_{\cdot j\cdot}-\bar{y}_{\cdots})^2, W_E = \sum_{ijk}(y_{ijk}-\bar{y}_{ij\cdot})^2 \\
W_{AB} = \sum_{ijk}(\bar{y}_{ij\cdot}-\bar{y}_{i\cdot\cdot}-\bar{y}_{\cdot j\cdot}+\bar{y}_{\cdots})^2,
\end{aligned}$$

where

$$\bar{y}_{\cdots} = \frac{1}{pqr}\sum_{ijk} y_{ijk}, \quad \bar{y}_{i\cdot\cdot} = \frac{1}{qr}\sum_{jk} y_{ijk}, \quad \bar{y}_{\cdot j\cdot} = \frac{1}{pr}\sum_{ik} y_{ijk}, \quad \bar{y}_{ij\cdot} = \frac{1}{r}\sum_{k} y_{ijk}\ .$$

As our competitor, the corresponding Bayes factor based on BIC is considered as follows.

$$\mathrm{BF}_{BIC}[\mathcal{M}_A;\mathcal{M}_1] = (pqr)^{-(p-1)/2}\left(1-\frac{W_A}{W_T}\right)^{-pqr/2}$$

$$\mathrm{BF}_{BIC}[\mathcal{M}_B;\mathcal{M}_1] = (pqr)^{-(q-1)/2}\left(1-\frac{W_B}{W_T}\right)^{-pqr/2}$$
$$\mathrm{BF}_{BIC}[\mathcal{M}_{A+B+1};\mathcal{M}_1] = (pqr)^{-(p+q-2)/2}\left(1-\frac{W_A+W_B}{W_T}\right)^{-pqr/2}$$
$$\mathrm{BF}_{BIC}[\mathcal{M}_{(A+1)(B+1)};\mathcal{M}_1] = (pqr)^{-(pq-1)/2}\left(\frac{W_E}{W_T}\right)^{-pqr/2}.$$

Now we give a result for model selection consistency of two-way ANOVA. The proof is omitted since it is straightforward in the very similar way as the proof of Theorem 3.1.

Theorem 4.1. *1. Assume $r \to \infty$ and p and q are fixed.*

(a) BF_{BIC} and BF_{FB} are consistent whichever the true model is.

2. Assume $p \to \infty$, $q \to \infty$ and r is fixed. Also let

$$\lim_{p\to\infty}\frac{\sum_i \alpha_i^2}{p\sigma^2} = c_A,\ \lim_{q\to\infty}\frac{\sum_j \beta_j^2}{q\sigma^2} = c_B,\ \lim_{p,q\to\infty}\frac{\sum_{ij}(\alpha\beta)_{ij}^2}{pq\sigma^2} = c_{AB}.$$

(a) BF_{BIC} and BF_{FB} are consistent whichever the true model is among $\mathcal{M}_1$, $\mathcal{M}_{A+1}$, $\mathcal{M}_{B+1}$, and $\mathcal{M}_{A+B+1}$.

(b) BF_{FB} is consistent under $\mathcal{M}_{(A+1)(B+1)}$ when

$$r^{1/(r-1)} < 1 + c_A + c_B + c_{AB} < (1+c_{AB})^r/r. \tag{4.2}$$

(c) BF_{FB} is inconsistent under $\mathcal{M}_{(A+1)(B+1)}$ when (4.2) is not satisfied.

(d) BF_{BIC} is inconsistent under $\mathcal{M}_{(A+1)(B+1)}$.

References

Berger, J. O., Ghosh, J. K. & Mukhopadhyay, N. (2003). Approximations and consistency of Bayes factors as model dimension grows. *J. Statist. Plann. Inference* **112**, 241–258.

Casella, G., Girón, F. J., Martínez, M. L. & Moreno, E. (2009). Consistency of Bayesian procedures for variable selection. *Ann. Statist.* **37**, 1207–1228.

George, E. I. & Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87**, 731–747.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90**, 773–795.

Liang, F., Paulo, R., Molina, G., Clyde, M. A. & Berger, J. O. (2008). Mixtures of g-priors for Bayesian variable selection. *J. Amer. Statist. Assoc.* **103**, 410–423.

Maruyama, Y. & George, E. I. (2008). A g-prior extension for $p > n$. arXiv:0801.4410v1 [stat.ME].

Maruyama, Y. & Strawderman, W. E. (2005). A new class of generalized Bayes minimax ridge regression estimators. *Ann. Statist.* **33**, 1753–1770.

Moreno, E., Girón, F. J. & Casella, G. (2009). Consistency of objective Bayes tests as the model dimension increases. Available at: `http://stat.wharton.upenn.edu/statweb/Conference/OBayes09/`.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.

Stone, M. (1979). Comments on model selection criteria of Akaike and Schwarz. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **41**, 276–278.

Strawderman, W. E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Statist.* **42**, 385–388.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian inference and decision techniques*, vol. 6 of *Stud. Bayesian Econometrics Statist.* Amsterdam: North-Holland, pp. 233–243.

Zellner, A. & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia (Spain)*, J. M. Bernardo, M. H. DeGroot, D. V. Lindley & A. F. M. Smith, eds. University of Valencia.